COMP 4432

Machine Learning

Group Project

Group 20

Student: Yu Ngo Ting Oscar

Student ID: 20048665D

Personal Contribution

- **What is your main role in doing this project?**

In this project, I am mainly responsible for several tasks, respectively planning the workflow of the model implementation, designing machine learning models, improving the model performance using different techniques, analyzing the model performance, and documenting.

As for the model implementation, I suggested to use six base model for initial observations, and our group would select the best one to implement the further improvement. In terms of the model improvement, I mainly focus on designing PCA and randomized search.

- **What other ideas you have considered to formulate the problem? Have your team members accepted or rejected them?**

Since our group has collected two datasets (fake headlines and real headlines), I have found out that the imbalance issue has existed in our combined dataset. While there are 50000 real headlines, there are only 11698 fake headlines. These issues make the base model performance too ideal, and the accuracy score becomes misleading high. To reduce the bias level, I suggest using the random sampling method to create a new balanced dataset. For this part, I collaborate with my groupmate Boris to learn to design the random sampling function.

- **What were the main difficulties encountered by the whole team? and by you individually?**

As for the main difficulties we encountered, we found out that the logistic regression did not obtain obvious improvement after we attempted different techniques (PCA, randomized search, grid search, and ensemble method).

To cope with this challenge, we decided to select one base model as our target to adopt the improvement techniques. Through discussion, we select kNN as the second model as target. Finally, it turns out that kNN could bring out a significant improvement.

- **What have you learned from this project?**

In this project, I have come to learn that the success of machine learning models lies in achieving mature model performance. However, this is greatly influenced by the characteristics of the dataset used. Therefore, it is critical to enhance the model architecture to obtain better improvement in performance.